ALIGNING SENTENCES IN BILINGUAL TEXTS
FRENCH - ENGLISH AND FRENCH - ARABIC

تقابل جمل لنصّين مختلفي اللغة
فرنسي-انكليزي وفرنسي-عربي

Fathi DEBILI, Elyès SAMMOUDA
CNRS-idl
Conseil d'Etat, Palais Royal
75001 Paris, France

فتحي الدبيلي - إلياس صمّودة
المركز القومي الفرنسي للبحث العلمي

ABSTRACT

In this paper, we will tackle the problem raised by the automatic alignment of sentences belonging to bilingual text pairs. The method that we advocate here is inspired by what a person with a fair knowledge of the other langage would do intuitively. It is based on the matching of the elements which are similar in both sentences. However, to match these elements correctly, we first have to match the sentences that contain them. There seems to be a vicious circle here. We will show how to break it. On the one hand, we will describe the hypotheses we made, and, on the other hand, the algorithms which ensued. The experiments are carried out with French-English and French-Arabic text pairs.

We will show that matching sentences and, later, expressions, amounts to raising a new problem in the machine translation field, i. e. the problem of *recognition* instead of that of translation, strictly speaking.

ملخّص

سنبحث في هذه المقالة مشكلة التقابل الآلي لجمل تنتمي إلى نصّين مختلفي اللّغة. الطريقة التي سنعتمدها هنا مقتبسة مما يقوم به الفرد الذي يعرف لغة أخرى معرفة متوسّطة ويرغب في إجراء هذا التقابل. هذه الطريقة تتركز على تقابل الكلمات المتماثلة في الجملتين المتناظرتين. ومن أجل تحقيق تماثل الكلمات هذا يجب أولا مقابلة الجمل التي تحتويها. قد تبدو هذه المسألة حلقة مفرغة إذ أن تماثل الكلمات يتطلّب تماثل الجمل وتقابل الجمل يتطلّب تماثل الكلمات. نبيّن في هذه المقالة كيفيّة الخروج من هذه الحلقة المفرغة ونشرح الفرضيات التي اعتمدناها وكذلك الخوارزميات الموافقة. التجارب التي تم إجراؤها اعتمدت على أزواج من النصوص: فرنسي - انكليزي وفرنسي - عربي.

سنبيّن أيضا أن مقابلة الجمل وفي المرحلة التالية مقابلة العبارات تطرح مشكلة جديدة في مجال الترجمة الآلية وهي مشكلة التعرف لا مشكلة الترجمة ذاتها.

REMERCIEMENTS : Le travail présenté a bénéficié de l'aide de nombreuses personnes. Nous les remercions toutes, en particulier *E. Souissi* et *A. Zribi* pour leurs contributions ; *E. Mackaay, I. Naddeo-Souriau, J.-L. Lemoigne* et la revue *Pour la Science* pour la gentillesse avec laquelle ils ont accepté de nous donner sur disquettes des textes ou fragments de textes monolingues ou bilingues ; *J. Kouloughli* et *J.-B. Berthelin* pour les discussions et critiques qu'ils sont toujours prêts à faire.

Cette recherche a été en partie financée par le Réseau des Industries de la Langue (contrat ACCT n° 338/SG/C5) et en partie par le MRT (décision d'aide n° 90.K.6434).

APPARIEMENT DES PHRASES DE TEXTES BILINGUES
FRANÇAIS - ANGLAIS ET FRANÇAIS - ARABES

Fathi DEBILI, Elyès SAMMOUDA*
CNRS-idl
Conseil d'Etat, Palais Royal - 75001 Paris, France
Tél.: (33-1) 43 50 54 01 ; Fax : (33-1) 40 20 83 70
* Systex, 91195 Saint Aubin, France

RESUME

Nous abordons dans ce papier le problème que pose la mise en correspondance automatique des phrases appartenant à des paires de textes bilingues. La méthode que nous préconisons s'inspire de ce que ferait intuitivement une personne connaissant moyennement l'autre langue. Elle se fonde sur l'appariement des éléments qui constituent les phrases en regard. Or, pour apparier correctement ces éléments, il faut au préalable avoir apparié les phrases qui les contiennent. Il y a là en apparence un cercle vicieux. Nous montrons comment le casser. Nous décrivons les hypothèses que nous faisons d'une part, et les algorithmes qui en découlent d'autre part. Les expérimentations sont effectuées sur les couples de langues *français-anglais* et *français-arabe*.

Nous montrons que l'appariement des phrases, et, dans l'étape d'après, des expressions, revient à poser un problème nouveau en traduction automatique : la *reconnaissance*, et non la traduction proprement dite.

Introduction :

Des gisements d'informations linguistiques

Ces dernières années ont vu apparaître plusieurs travaux qui prennent pour départ des corpus bilingues. L'intérêt grandissant porté à ces corpus est sans doute lié à leur disponibilité sur supports informatiques. Mais là n'est pas la raison essentielle. Celle-ci réside plutôt dans l'idée selon laquelle ces corpus sont autant de *mines*, encore très largement inexploitées, qui renferment une très grande quantité d'informations fort utiles pour les diverses problématiques que pose la traduction, que celle-ci soit automatique ou non.

Des objectifs divers

Les perspectives d'exploitation de ces *gisements* sont diverses. Pour P. Brown *et al.* d'IBM, le but est de calculer les paramètres du modèle probabiliste de traduction automatique qu'ils veulent construire [Brown *et al.*, 1988 ; Brown *et al.*, 1990]. Pour [Catizone *et al.* 1989], c'est fournir des outils aux lexicographes et aux linguistes qui étudient la traduction. Pour [Sadler, 1989], l'idée est de construire une banque de connaissances bilingues. [Sato et Nagao, 1990] posent quant à eux le problème de la traduction par l'exemple.

Notre objectif

Il est de construire automatiquement des dictionnaires de transfert d'expressions. En partant de paires de textes analysés sous forme de dépendances, nous essayons de mettre en correspondance des morceaux d'arbres [Debili, 1990].

La reconnaissance : un problème nouveau

Le problème de la traduction n'est donc pas abordé sous l'angle traditionnel de l'*analyse-synthèse*. Il est considéré sous l'angle de la *reconnaissance*. La différence est importante, puisqu'il s'agit dans un cas de traduire, et dans l'autre, de reconnaître qu'il y a bien traduction. Dans le premier cas, l'entrée du système est un texte, et la sortie, sa traduction. Dans le second cas, l'entrée est un couple de textes pris dans deux langues, et la sortie est la réponse à la question : *sont-ils traduction l'un de l'autre ?* et si oui, *dans quelle mesure ?*

A partir de là, les applications sont, on l'imagine, nombreuses. Elles vont de l'évaluation des systèmes de traduction automatique à la détection du plagiat.

Un dénominateur commun

Tous ces travaux ont en commun d'essayer d'extraire à partir de paires de textes bilingues, divers types d'informations linguistico-statistiques, et en cela, de contribuer aux efforts qui sont faits notamment dans le domaine de la traduction automatique. Ils diffèrent par les objectifs, et surtout par les méthodes et les moyens mis en oeuvre.

L'appariement

Un sous-problème commun

Un point de passage obligé dans tous ces projets est la nécessaire mise en correspondance des éléments (phrases, syntagmes, mots, etc.) qui composent les textes en regard. C'est cet aspect du problème, que la terminologie anglaise semble déjà consacrer sous le nom d'*alignement*, que nous abordons dans le présent papier.

La solution informatique n'est pas simple, même si, à l'exercice, la tâche se révèle relativement aisée pour qui connaît moyennement l'autre langue. Comme chacun sait en effet, le processus de traduction, même s'il essaie de se conformer au principe de la conservation de la succession des "idées", ne maintient pas toujours, au plan de la réalisation textuelle, le même découpage en paragraphes, phrases, syntagmes, etc.. Le nombre et la taille de ces unités n'est donc pas le même dans les textes en regard, ni non plus leur ordre, ni, loin s'en faut, leur forme. C'est le contraire qui constitue l'exception.

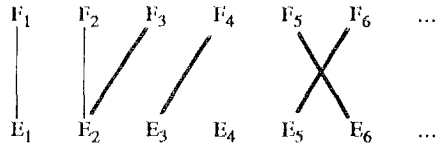
Des approches différentes

Quatre équipes, à notre connaissance, se sont intéressées au problème. Deux méthodes semblent se dégager. L'une, *statistique*, ne se fonde que sur les caractéristiques formelles des textes en regard: c'est la taille des phrases exprimée en nombre de mots pour [Brown *et al.*, 1991] ; c'est la taille des phrases exprimée en nombre de caractères pour [Gale et Church, 1991]. L'autre, *linguistique* dirons-nous, s'inspire de ce que nous ferions nous-mêmes intuitivement : elle se fonde sur l'appariement des unités lexicales qui composent les phrases, accompagnées éventuellement de leur structure dépendancielle. Elle fait intervenir en quelque sorte le contenu. C'est la voie qui est suivie par Catizone *et al.*, et par M. Kay, cité dans [Catizone *et al.*, 1989]. C'est aussi celle que nous avons empruntée. Il reste que dans les deux cas on fait aussi intervenir, quand on en dispose, différents marqueurs, comme par exemple les délimiteurs de sections, de chapitres, de paragraphes, etc.. Leur mise à profit améliore certes les résultats, mais elle n'est pas obligatoire.

Les deux méthodes ne s'opposent pas. Elles se complètent plutôt. Leurs potentialités sont en effet différentes. Pour l'essentiel, l'approche statistique convient mieux si les corpus sont très gros, car elle est plus rapide. Si, par contre, les corpus sont de faible taille, à l'extrême, si l'on veut simplement savoir dans quelle mesure deux phrases données sont traduction l'une de l'autre, alors c'est incontestablement l'approche linguistique qui l'emporte, l'autre pouvant même devenir impraticable.

L'appariement des phrases

Il consiste à jumeler les phrases qui se correspondent dans un couple de textes bilingues ; c'est-à-dire à relier les phrases qui sont traduction les unes des autres. Le problème vient de ce qu'il n'y a pas toujours correspondance biunivoque entre les phrases des textes en regard. A une phrase il peut en correspondre deux, parfois plus. Dans d'autres cas, il n'en correspondra aucune. En outre, il arrive qu'il y ait inversion. La figure suivante illustre ces différents cas. F_i désigne les phrases du texte français, E_j celles du texte anglais.



D'autres cas plus complexes encore peuvent se produire, même s'ils restent rares. Par exemple ceux de la figure suivante :

Pourquoi apparier des phrases ?

Pourquoi faut-il apparier les phrases ? Pourquoi ne pas tenter d'apparier directement les morceaux de structures dépendancielles correspondant aux expressions, puisque c'est là le but final recherché ? Parce qu'il est plus facile d'*approcher puis* d'*atteindre* que d'atteindre d'emblée.

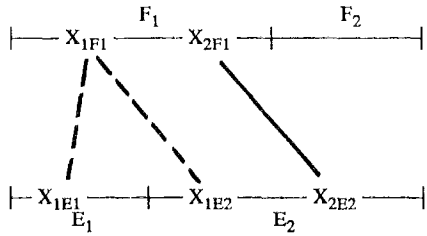
A l'appui, plusieurs raisons interdépendantes.

1.- Réduire la combinatoire : l'idée est de délimiter d'abord les champs d'investigation, c'est-à-dire les parcelles du texte en vis-à-vis où est susceptible de se trouver l'expression correspondante recherchée. Or les expressions que nous voulons mettre en correspondance sont internes à des phrases. Nous aimerions donc que la recherche se fasse à l'intérieur de l'unité que constitue la phrase. De sorte que, les possibilités de choix étant réduites, les probabilités d'erreurs s'en trouvent à leur tour minimisées.

2.- Utiliser une règle de cohésion : à rechercher les correspondances dans une paire de phrases, on fait intervenir les contextes dans

lesquels baignent les expressions que nous souhaitons apparier. En cas d'ambiguïté, l'appariement des éléments alentours, s'il n'est pas ambigu, aidera à faire la décision. Nous pouvons penser que nous aurons ainsi moins d'erreurs. *L'hypothèse sous-jacente que nous faisons est que les constituants d'une phrase ont tendance à rester ensemble lors du processus de traduction.*

Dans la figure suivante, X_{1F1} désigne l'expression de rang 1 dans la phrase française F_1, etc..



La règle de cohésion stipule que dans le cas présent X_{1F1} doit être mis en correspondance avec X_{1E2} de la phrase anglaise E_2, en raison de la présence du lien qui existe entre X_{2F1} et X_{2E2}, plutôt qu'avec X_{1E1} de la phrase E_1. L'ambiguïté est donc ainsi levée. Cette règle n'est cependant pas absolue.

3.- Ces raisons sont confortées par le fait qu'apparier des phrases semble a priori plus simple qu'apparier des expressions. En effet leur délimitation est plus simple, il y a donc moins d'erreurs. Par ailleurs, étant plus longues, elles présentent davantage de points d'ancrage pour la mise en correspondance. Les risques d'erreur sont donc là aussi moindres. Enfin, étant moins nombreuses, les possiblités de choix seront réduites d'autant, et les probabilités d'erreurs aussi.

4.- Conservation de la séquentialité des "idées" dans le processus de traduction :

Enfin, en faveur de l'appariement des phrases d'abord, puis des expressions ensuite (que nous avons résumé par la formule *approcher puis atteindre*), il y a l'hypothèse que nous faisons de la *conservation de la séquentialité des "idées" dans le processus de traduction.* Le corollaire de cette hypothèse peut être présenté comme suit :

> *dans le processus de traduction,*
> ***la séquentialité des chapitres d'un livre par exemple est davantage respectée que***
> ***la séquentialité des paragraphes qui le constituent, laquelle est davantage respectée que***
> ***celle des phrases, plus respectée que***
> ***celle des propositions, elle-même mieux conservée que***
> ***celle des expressions, qui, à son tour, est mieux conservée que***
> ***celle des mots, dont nous ajoutons, si cela a un sens, qu'elle est elle-même bien entendu mieux conservée que***
> ***celle des caractères.***

Il y a en quelque sorte de plus en plus de désordre à mesure que l'on descend dans l'échelle. Dans cette liste, une cassure semble a priori se produire au niveau du point d'entrée "phrases". En effet, au delà, on est sûr que l'ordre sera souvent altéré. La phrase est au prime abord la plus petite unité dont l'ordre sera presque toujours maintenu.

En conséquence, la séquentialité des phrases étant davantage respectée que celle des expressions, nous ferons en principe moins d'erreurs à apparier les phrases d'abord, puis les expressions ensuite, qui en bénéficieront.

L'appariement des phrases apparaît donc clairement comme une étape préalable aux diverses autres étapes d'appariements que nous pourrions imaginer (celles des expressions nominales, verbales, etc.).

Nos hypothèses

Dans ce paragraphe, nous donnons les hypothèses qui sont à la base de notre algorithme. Leur mise en oeuvre doit en principe conduire au meilleur appariement possible entre phrases. Elles se traduiront pour la plupart par des critères de choix.

1.- Tout d'abord, nous ne tenons aucunement compte de la distinction *source-cible*, attachée pourtant de fait aux paires de textes retenues. Pour un couple de langues données, il se pourrait en effet que les paires de textes aient des propriétés formelles différentes selon le sens de traduction. Par exemple, le rapport du nombre moyen de mots de la phrase française au nombre moyen de mots de la phrase anglaise pourrait être différent selon que les textes ont été traduits dans un sens ou dans l'autre. Nous n'avons pas fait de mesures dans ce sens. Quoiqu'il en soit, nous supposons pour l'instant que cette distinction n'est pas déterminante pour l'appariement des phrases. Autrement dit, nous faisons l'hypothèse de la bidirectionnalité des appariements que nous souhaitons reconnaître.

2.- Nous disons que deux phrases sont d'autant plus traduction l'une de l'autre que :

i) elles présentent les "mêmes" mots,

ii) en outre, ces mots entretiennent entre eux des relations analogues.

L'appariement des phrases repose donc sur l'appariement des mots. Or, dans le paragraphe précédent, nous avons expliqué que pour apparier des unités plus petites, en l'occurrence des mots, il fallait auparavant avoir apparié les phrases qui les contiennent. En résumé :

pour apparier les mots il faut apparier les phrases
pour apparier les phrases il faut apparier les mots

Nous sommes donc en présence d'un cercle vicieux. Pour le briser, il faut en fait dire simplement :

Pour obtenir un appariement fin des mots il faut apparier les phrases ;
Pour apparier les phrases on peut se contenter d'un appariement grossier des mots.

3.- Proximité de rang

Le principe de la conservation de la séquentialité des phrases, même s'il n'est pas toujours respecté, nous amène à faire l'hypothèse que *les rangs des phrases en correspondance sont en général proches*. C'est-à-dire que si F_i et E_j sont traduction l'une de l'autre, alors $|i-j|$ est toujours inférieur à un certain seuil s dont la valeur pourrait être déterminée à partir de l'observation des textes à aligner. Nous avons fixé a priori ce seuil à 3. Autrement dit, le correspondant de chaque phrase est supposé être dans une fenêtre du texte en vis-à-vis dont la largeur est égale à $2xs + 1$, soit 7 phrases dans le cas présent.

Attention ! i et j ne sont pas des rangs absolus. Ils sont relatifs. Il faut en effet rattraper les décalages introduits par les appariements autres que (1, 1). i et j sont donc réinitialisés régulièrement, au fur et à mesure que l'on avance dans l'appariement.

Cela veut dire aussi qu'en cas d'ambiguïté, ce sont les phrases de rang proche qui seront privilégiées. Au plan algorithmique, ceci sera obtenu en multipliant la note mesurant la parenté de deux phrases par le facteur suivant :

$\alpha = 1 - | i - j | / (i + j) \qquad 0 < \alpha <= 1$

4.- Proximité de taille

En relation avec le point 4 de la page précédente, nous supposons que *le processus de traduction conserve davantage l'égalité :*

***du nombre de chapitres** d'un livre par exemple*
***que celui des paragraphes** par chapitre*
***que celui des phrases** par paragraphe*
que celui (dans l'ordre) des
propositions
expressions
mots pleins
ou mots vides
par phrase.

C'est là une expression généralisée de l'idée simple selon laquelle une phrase aura tendance à être traduite par une phrase longue si elle est longue, et par une phrase courte si elle est courte.

En cas d'ambiguïté, nous chercherons par conséquent à privilégier les phrases de taille voisine.

La mise en oeuvre de cette hypothèse est pour l'instant simple. Elle consiste à faire intervenir un coefficient de normalisation défini comme suit :

$ß = 1 - | L(F_i) - L(E_j) | / (L(F_i) + L(E_j))$
$0 < ß <= 1$

dans le calcul de la distance entre deux phrases.

L(P) est la taille de la phrase P exprimée en nombre de mots.

Défini ainsi, ce facteur ne tient pas compte du rapport pratiquement constant qui s'établit entre les tailles moyennes des phrases pour un couple de langues données. Pour le couple français-anglais par exemple, ce rapport est d'environ 1,15. C'est-à-dire que les phrases françaises sont en moyenne plus longues d'environ 15%. L'utilisation de ß tel quel conduirait donc à favoriser des phrases de même taille, alors qu'il faudrait favoriser celles dont la taille est légèrement plus longue ou plus courte, selon la langue de départ.

Pour rattraper cette différence de taille, une solution simple consiste à faire intervenir ce rapport dans la définition de ß. Il suffit de multiplier la taille de l'une ou de l'autre phrase par un facteur correctif, - dans le cas présent, la taille de la phrase anglaise par 1,15.

Nous avons songé un moment à ne retenir dans le calcul de la taille des phrases que les mots pleins, et non la totalité. Cet espoir s'est en fait très vite envolé. Le rapport des tailles moyennes des phrases française et anglaise ne baisse que de quelques centièmes lorsque l'on ne comptabilise que les mots pleins.

L'algorithme

Pour simplifier l'exposé nous allons considérer le cas français-anglais en adoptant les notations suivantes :

F_I :	est la I-ème phrase du texte français
E_J :	est la J-ème phrase du texte anglais
f_i :	est la forme canonique du i-ème mot de F_I
e_j :	est la forme canonique du j-ème mot de E_J
$e_{k,i}$:	est la k-ième traduction anglaise de f_i
$f_{l,j}$:	est la l-ième traduction française de e_j

En outre, nous n'allons considérer pour l'instant que les cas d'appariements (1, 1). Nous verrons par la suite comment traiter les autres cas: (1, 0), (1, n), et (n, m) avec n et m > 1.

L'appariement des phrases

Le problème est de reconnaître pour une phrase donnée du texte de départ, disons F_I, la phrase qui lui correspond le mieux dans le texte d'arrivée, disons E_J. Mais attention, cela ne suffit pas pour décréter l'appariement (F_I, E_J). Il faut en effet qu'il y ait réciprocité : F_I doit être à son tour la meilleure phrase candidate pour la phrase de départ E_J.

La figure suivante illustre ce que le programme doit effectuer.

Chacune des phrases des deux textes doit être comparée à l'ensemble des phrases qui sont susceptibles de lui correspondre. L'appariement se fera ensuite sur la base de ces comparaisons.

Dans le cas présent, F_I doit être comparée à toutes les phrases de la fenêtre censée contenir la phrase E_J recherchée. Inversement, E_J doit être comparée à toutes les phrases de la fenêtre qui lui est associée. Si F_I et E_J sont mutuellement meilleure traduction l'une de l'autre, alors leur appariement est retenu. Cette condition est très restrictive. Nous verrons les modulations qui lui doivent être apportées pour construire les appariements autres que (1, 1).

La comparaison de deux phrases

Elle repose sur l'appariement des mots qui les composent. Plus cet appariement est dense, plus les phrases sont proches ; moins il est dense, moins elles sont proches. Nous cherchons à calculer une note qui puisse refléter cette proximité. Nous voudrions qu'elle soit d'autant plus importante que :

i) les deux phrases comportent les mêmes mots,
ii) que ces mots sont longs,
iii) et que leur séquentialité est respectée.

Les relations de dépendance n'interviennent donc pas. C'est que nous supposons possible de s'en passer pour l'instant, et que pour l'appariement des phrases, l'on peut par conséquent se contenter d'un appariement relativement grossier des mots.

L'appariement des mots

A bien des égards, il est analogue à celui des phrases. Les problèmes sont similaires ; les solutions semblables. En particulier, les hypothèses de rang et de taille sont transposables.

Considérons deux phrases F_I et E_J. L'appariement des mots qui les composent est obtenu en comparant successivement chacun des mots de F_I à tous les mots de E_J. Les comparaisons (f_i, e_j) sont établies à l'aide d'un dictionnaire de transfert de mots simples. Les résultats sont consignés dans une matrice - *Matmot* - dont les lignes correspondent aux mots de F_I, et les colonnes, aux mots de E_J.

La comparaison de deux mots

Chaque élément (f_i, e_j) de la matrice *Matmot* reçoit une note que nous voulons d'autant plus forte que les deux mots f_i et e_j sont traduction l'un de l'autre. Nous voulons en outre que cette note reflète la taille des mots comparés ; et qu'enfin, elle tienne compte de la proximité de leurs rangs respectifs dans les phrases d'où ils sont extraits.

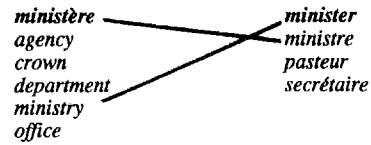
Le calcul de la note s'établit de la façon suivante. Les traductions $e_{k,i}$ de f_i sont comparées une à une à e_j, et celles de e_j, c'est-à-dire $f_{l,j}$, à f_i. Se pose par conséquent un problème devenu classique : celui de la *comparaison dynamique*. Nous utilisons l'algorithme de Bellman [Bellman, 1957 ; Laurière, 1979] non pour calculer une distance, mais pour déterminer les sous-chaînes maximales communes aux deux graphies comparées. Le problème n'est pas vu sous l'angle du coût à payer pour passer de l'une à l'autre. Il est plutôt de considérer que deux graphies sont d'autant plus proches qu'elles ont en commun les mêmes sous-chaînes, et que celles-ci sont les plus longues possibles. La note que nous avons retenue pour l'instant, et qui s'est montrée satisfaisante dans une autre problématique, est donnée par la formule suivante : elle est égale à la somme des carrés des tailles des sous-chaînes maximales communes multipliée par un facteur de normalisation analogue à ß.

$$N = [1-(|L(c_1)-L(c_2)|/(L(c_1)-L(c_2)))]\Sigma n(t).t^2$$

où L(c) est la taille en nombre de caractères de la chaîne c, et n(t) le nombre de sous-chaînes maximales communes de longueur t.

Chaque comparaison se traduit par une note. Nous notons $N_{f\text{-}e}$ la meilleure note obtenue dans le sens français-anglais, et $N_{e\text{-}f}$ dans le sens contraire. La note globale est obtenue en additionnant $N_{f\text{-}e}$ et $N_{e\text{-}f}$. Cette note est dite intrinsèque car elle ne tient pas compte des rangs des deux mots dans leurs phrases respectives. Afin de favoriser précisément les mots de rang proche, le résultat est multilplié par un facteur analogue à α.

Prenons un exemple :



$N_{f\text{-}e}=[1-(|9-8|/(9+8))](6^2+2^2)=37{,}647$
$N_{e\text{-}f}=[1-(|8-8|/(8+8))](6^2+1^2)=37$

Pour être plus discriminantes encore, les notes partielles $N_{f\text{-}e}$ et $N_{e\text{-}f}$ pourraient être calculées en faisant intervenir d'autres données contextuelles comme par exemple les catégories grammaticales ou du nombre des graphies comparées.

Retour à l'appariement des mots

Pour une paire de phrases F_I et E_J données, nous disposons maintenant d'une matrice remplie. Il importe de remarquer que l'appariement des mots ne peut être recherché à cette étape du traitement. En effet, nous ne sommes pas encore sûrs que F_I et E_J sont bien traduction l'une de l'autre. Les appariements que nous allons faire sont donc hypothétiques. Ils ne servent qu'à marquer les éléments de la matrice qui interviendront dans la comparaison de deux phrases.

L'appariement des mots est obtenu en effectuant un double balayage de la matrice. Pour chaque ligne f_i on détermine la meilleure colonne e_j, que nous marquons à l'aide de la lettre ***e***. Nous indiquons ainsi que le mot e_j est, parmi les mots de la phrase E_J, la meilleure traduction de f_i ; sa note étant baptisée $T_{f\text{-}e}$. Nous faisons de même pour chaque colonne, mais nous marquons cette fois d'un ***f*** la meilleure ligne retenue, la note correspondante étant baptisée $T_{e\text{-}f}$. Lorsqu'il y a coïncidence nous mettons ***x*** pour indiquer que les deux mots sont mutuellement meilleure traduction l'un de l'autre.

Ainsi construite, la matrice *Matmot* représente l'ensemble des liens qui s'établissent entre deux phrases données. Ces liens sont orientés. Ils sont en outre caractérisés par un nombre censé mesurer leur force.

Retour à la comparaison de deux phrases

Le but est de mesurer la force globale de ces liens. Il s'agit par conséquent de traduire la matrice qui a servi à les établir en un scalaire. La solution retenue est simple : elle consiste à calculer deux notes partielles. La première, $N_{F\text{-}E}$, est obtenue en additionnant les valeurs maximales rencontrées en parcourant les lignes de la matrice. La deuxième, $N_{E\text{-}F}$, est obtenue en parcourant les colonnes. Les deux notes étant bien entendu multipliées par le facteur de normalisation ß.

$$N_{F\text{-}E} = ß \sum_i \max_j Matmot(f_i, e_j) = ß \Sigma T_{f\text{-}e}$$

$$N_{E\text{-}F} = ß \sum_j \max_i Matmot(f_i, e_j) = ß \Sigma T_{e\text{-}f}$$

La première note reflète en quelque sorte la force avec laquelle E_J est traduction de F_I. La seconde, la force avec laquelle F_I est traduction de E_J. La note globale est obtenue en additionnant ces deux notes partielles. Nous disons que cette note est intrinsèque car elle ne fait pas encore intervenir les rangs dans leurs textes respectifs des deux phrases comparées.

Le facteur multiplicatif qui joue ce rôle, c'est-à-dire α, intervient en fait lors de l'appariement des phrases. C'est que les rangs considérés ne peuvent être des rangs absolus. En effet, pour des textes longs, des décalages importants peuvent s'introduire, suite aux appariements autres que (1, 1). Ne pas en tenir compte conduirait à défavoriser des paires de phrases qui pourtant sont bien traduction l'une de l'autre. L'idée est donc de considérer les rangs qu'ont les phrases dès lors qu'elles entrent dans l'une des deux fenêtres de travail. Plus encore, α sera déterminé de façon dynamique, puisque les positions relatives des deux phrases considérées changeront au fur et à mesure que seront évacuées les phrases qui les précèdent.

Remarque : Quoique la comparaison de deux phrases puisse rappeler la comparaison de deux chaînes de caractères, comme par exemple dans le domaine de la détection-correction des graphies fautives, ou de deux images acoustiques, comme en reconnaissance de la parole, on ne peut en adopter la solution de la comparaison dynamique.

Il est difficile en effet d'imaginer ce que pourrait être le calcul d'une distance considérée comme étant le coût des opérations à effectuer pour passer d'une phrase à une autre. Il y a en outre une différence fondamentale : alors que dans le premier cas, la séquentialité parallèle des événements à comparer est une donnée, dans le second, cette séquentialité n'est nullement une donnée du problème. Dans le premier cas, il s'agit d'ajuster des événements qui ont des indices différents mais qui néanmoins se présentent selon la même séquentialité, ou presque. Dans le second, les éléments des deux phrases peuvent se correspondre indépendamment de leur séquentialité respective. Par exemple, le premier mot de l'une des deux phrases peut très bien correspondre au dernier mot de l'autre.

Retour à l'appariement des phrases

L'appariement des phrases de deux textes ressemble à l'appariement des mots de deux phrases. La construction des appariements repose sur l'emploi d'une matrice dont les lignes correspondent aux phrases françaises, et les colonnes, aux phrases anglaises. Chaque élément de la matrice reçoit la note issue de la comparaison des deux phrases correspondantes. Si

deux phrases se déplacent dans la matrice, alors leur note est réévaluée puisque leurs rangs relatifs ont changé. L'appariement des phrases est fondé sur la détermination des notes maximales lignes et colonnes. Seules les phrases qui occupent dans le cas présent les quatre premières lignes ou colonnes peuvent être appariées, et par suite évacuées de la matrice. Après quoi il y a décalage, puis entrée des phrases suivantes, et ainsi de suite.

Résultats

L'algorithme, qui n'est pas encore figé, donne dans sa version d'aujourd'hui des résultats qui vont du médiocre à l'excellent, selon la nature des textes soumis. Sur les deux textes juridique et technique qui ont servi aux expérimentations, les résultats sont satisfaisants comme le montre le tableau partiel suivant qui en donne le cumul.

Nb phr. fr.:339 ; Nb phr. angl. : 350

	(1, 1)		(1, 2)		(1, 3)	
(1,1)	269	83.5%	34	10.5%	4	1,2%
	12	3,7%	1	0.3%	0	0.0%
(1,2)	5	33,3%	8	53,3%	0	0.0%
	0	0.0%	2	13.3%	0	0.0%

où l'on doit lire que parmi les relations (1,2) par exemple, 8 ont été reconnues, deux correspondent à du bruit, et 5 sont partiellement reconnues.

Sur les paires de textes provenant de la revue *Pour la Science*, les résulats sont plutôt médiocres. C'est que dans ce cas il y a "réécriture" plutôt que traduction. Il y a aussi que les taux de couverture en usage des dictionnaires de transfert sont relativement faibles, de l'ordre de 57 %, parfois moins, selon les textes.

Nous ne pouvons détailler et donner dans le peu de place qui reste les différents tableaux de résultats que nous avons obtenus. L'idée générale qui se dégage est que l'algorithme est au stade actuel davantage bruyant que silencieux. Enfin, il convient de souligner que les relations de type (1,0) ne sont pour le moment nullement reconnues.

C'est donc dire combien le problème de l'appariement des phrases, qui selon nous pose celui de la reconnaissance de la traduction, est difficile.

Conclusion

L'algorithme que nous avons présenté est relativement simple. Il repose sur la construction de deux matrices. La première permet de comparer deux phrases en en appariant les mots. La seconde permet, en comparant deux textes, d'en proposer les meilleurs appariements de phrases. On considérera qu'il est peut être coûteux en temps. C'est là un aspect que nous avons délaissé volontairement, voulant pour l'instant réussir à reconnaître si deux phrases données sont bien traduction l'une de l'autre, et si oui, dans quelle mesure. - D'abord dans des conditions facilitant grandement la tâche, car les phrases sont extraites de textes dont on sait a priori qu'ils sont traduction les uns des autres. En ce sens, la reconnaissance est contextuelle, car elle est fondée davantage sur le rejet des phrases alentours qui ne vont pas, que sur une véritable reconnaissance de celle(s) qui vont. - Puis dans des conditions plus difficiles, en essayant de répondre sans l'appui de cette connaissance apriorique.

Bibliographie

[Bellman, 1957] Bellman, R. *Dynamic Programming*, Princeton University Press.

[Brown *et al.*, 1988] Brown, P., J. Cocke, S. Delle Pietra, V. Delle Pietra, F. Jelinek, R. Mercer, and P. Roossin. "A Statistical Approach to Language Translation", In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary (1988).

[Brown *et al.*, 1990] Brown, P., J. Cocke, S. Delle Pietra, V. Delle Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. "A Statistical Approach to Machine Translation", In *Computational Linguistics*, Vol. 16, Number 2, pp. 79-85 (June 1990).

[Brown *et al.*, 1991] Brown, P., J. Lai, R. Mercer. "Aligning Sentences in Parallel Corpora", In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, (1991).

[Catizone *et al.*, 1989] Catizone, R., G. Russell, and S. Warwick "Deriving Translation Data from Bilingual Texts", In *U. Zernick (ed.) Proceedings of the First Lexical Acquisition Workshop*, Detroit, (1989).

[Debili, 1990] Debili, F. "Construction automatique de dictionnaires de transfert d'expressions français-anglais et français-arabe". *Rapport final d'exécution du Projet soumis au Réseau Français des Industries de la Langue, Contrat n° 338/SG/C5, ACCT-UA 962 du CNRS*, Paris (31/12/1990).

[Gale and Church, 1991] Gale, W. and K. Church "A Program for Aligning Sentences in Bilingual Corpora", In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, (1991).

[Laurière, 1979] Laurière J. L. *Eléments de programmation dynamique*, Gauthiers-Villars.

[Sadler, 1989] Sadler V. "The bilingual Knowledge Bank - A new conceptual basis for MT". *BSO/Research*, Utrecht (1989).

[Sato and Nagao, 1990] Sato, S. and M. Nagao. " Toward Memory-based Translation", In *Proceedings of the 13th Int. Conference on Computational Linguistics*, Helsinki, (1990).

[Warwick and Russell, 1990] Warwick S. and G. Russell, "Bilingual Concordancing and Bilingual Lexicography", In *EURALEX 4th International Congress*, Malaga, Spain (1990)